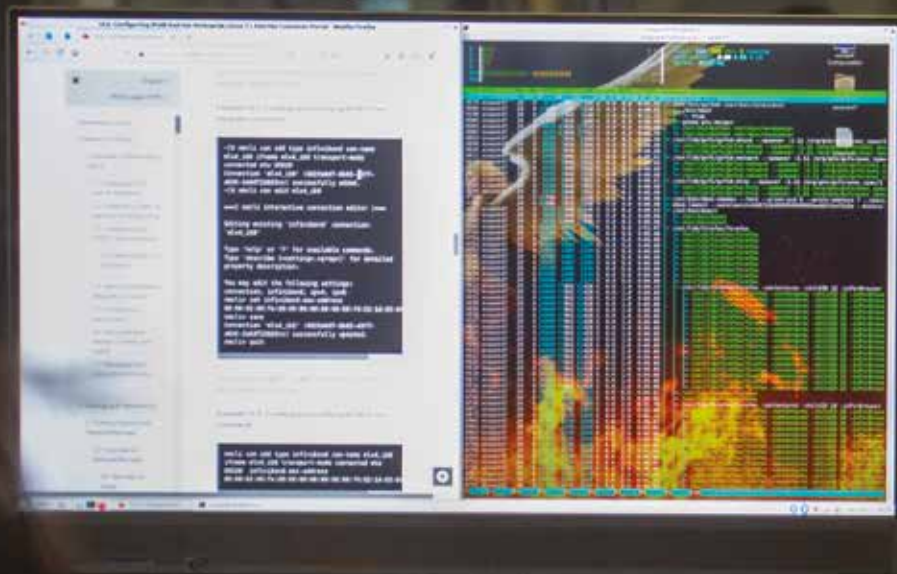


Caoba y los cazadores de datos en EAFIT



Explorar el **big data** para tomar mejores decisiones en empresas y el Estado es uno de los objetivos del Centro de Excelencia Caoba que se especializa en extraer información valiosa de los datos a gran escala para analizarlos y generar valor agregado para múltiples usos.

SANTIAGO HIGUITA POSADA
Colaborador

Caoba ha logrado reconocimiento nacional como líder en *big data* y *data analytics*. Producto de esta búsqueda surgió la maestría en Ciencias de Datos y Analítica de EAFIT. Foto Robinson Henao.

En medio de la avalancha de datos que se producen todos los días en internet, una de las mayores unidades para el almacenamiento de información hasta ahora conocidas es el Zettabyte (ZB). Equivale a mil trillones de bits, la unidad mínima empleada en informática. Su capacidad es tal que en un ZB se podrían alojar más de 17 mil millones de iPhone de 64 gigas y una película en alta definición duraría cerca de 36 millones de años.

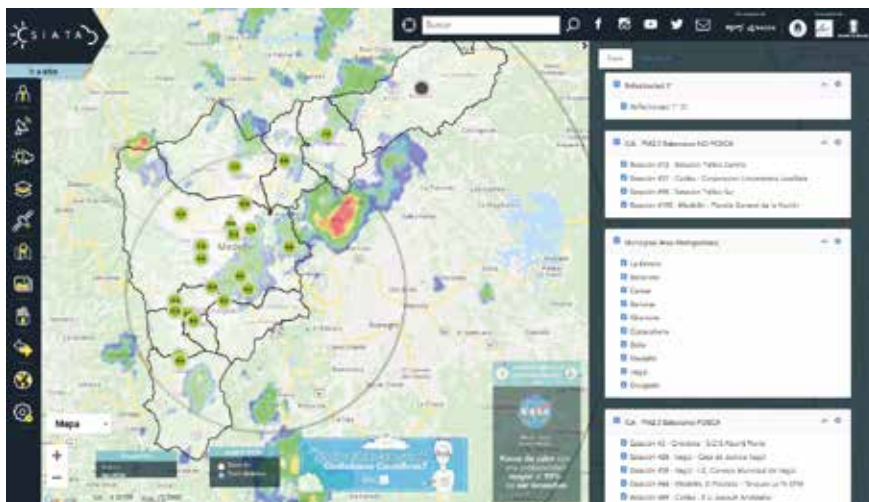
Para el año 2021 se pronostica que el tráfico de los centros de datos en el planeta, gran parte almacenados en la nube computacional, podría superar los 20 ZB anuales, más del doble del peso actual de la red mundial. Firms internacionales como Cisco, líder en transformación digital

y estudio de *big data*, han visualizado este panorama en la producción de información en su *Índice Global de la Nube*.

En este mar de datos, literalmente, navegan los *data scientist* que son profesionales en áreas convergentes como matemática, modelación, estadística y computación dedicados a auscultar la inmensidad de datos producidos a diario. Por medio de técnicas de *data mining* hallan información valiosa que sirve para tomar mejores decisiones tanto en el sector privado como en políticas públicas.

Son los científicos de la era de la información que se dedican a pensar, principalmente desde la ingeniería y la abstracción matemática, aplicaciones para hurgar en los recovecos del *big data* y crear algoritmos que

Solo el 65 % de los datos producidos en Colombia es recolectado. De ellos, 58 % es relevante, según el Ministerio de Tecnologías de la Información y las Comunicaciones.



Información de satélites y estaciones de monitoreo permiten al Siata (Sistema de Alerta Temprana del Valle del Aburrá) el seguimiento de diversos fenómenos hidrometeorológicos en esta zona y sus alrededores.

solucionen problemas que afectan la cotidianidad desde el sector financiero hasta el medioambiente.

Edwin Montoya Múnera es ingeniero de sistemas y doctor en Telecomunicaciones de la Universidad Politécnica de Valencia (España). En la actualidad es profesor investigador de EAFIT y coordinador del Centro de Excelencia y Apropiación en *Big Data* y *Data Analytics* de la Alianza Caoba, una iniciativa público-privada impulsada por el Gobierno Nacional para promover investigación aplicada en estas áreas. La iniciativa opera desde el año 2016 en el cuarto piso del bloque 18 de EAFIT.

"La ciencia de datos necesita de dos grandes insumos: uno son los grandes datos, lo segundo es la infraestructura para procesarlos. La Universidad cuenta con infraestructura tecnológica de alto nivel, desde el supercomputador Apolo hasta servidores intermediarios en el Laboratorio de Sistemas, que hemos usado para procesar los datos masivos", comenta Edwin Montoya, quien es además jefe del Departamento de Ingeniería de Sistemas de la Universidad y líder de Caoba.

LOS DATOS, EL NUEVO PETRÓLEO

La minería de datos, una de las técnicas para la extracción de información y conocimiento a partir del *big data*, se popularizó durante la década de 1990 con la masificación de las redes informáticas. El término se refiere al *data analytics* fundamentado en métodos estadísticos y de aprendizaje computacional como el *machine learning*, técnicas que se perfeccionan con el avance tecnológico.

Para el coordinador del Centro de Excelencia Caoba, durante los cinco años en que se ha establecido la plataforma se han generado distintas soluciones mediante la investigación aplicada, consultoría, formación y transferencia de conocimiento en estos campos, contribuyendo a la competitividad del país.

PROYECTOS DEL CENTRO DE EXCELENCIA

El Departamento Nacional de Planeación, Bancolombia y la empresa Nutresa fueron instituciones que implementaron la analítica de datos en sus organizaciones mediante las investigaciones de Caoba. Algunas de ellas fueron:



Comunidades financieras:

Se realizó un análisis para determinar las redes de interacción comercial y pronosticar el riesgo de los usuarios con la implementación de modelos econométricos y algoritmos.



Calidad del servicio de agua potable:

Adelantado por el Departamento Nacional de Planeación. Su objetivo fue desarrollar una herramienta para el análisis de redes y cuantificar de forma diferenciada las inversiones públicas en el servicio de acueducto en los municipios de Colombia.



Sistema de integración y análisis para medios digitales de Siamed Nutresa:

Proyecto de analítica descriptiva y predictiva aplicado al mercadeo de la compañía mediante el monitoreo de sus consumidores.

"El análisis de datos es una tecnología relativamente vieja, es decir, tiene más de 20 años en los cuales las empresas ya hacían analítica", recuerda Montoya. "Las fuentes de datos eran los mismos datos de las compañías y no se habían desarrollado modelos tan sofisticados de computación que permitieran sacarles más provecho. Con el aumento del procesamiento y la mayor diversidad de generación hubo una explosión de datos", complementa.

Este fenómeno ha venido creciendo de manera exponencial en la última década. En la actualidad, un gran volumen de *big data* se produce a diario de manera frenética: correos electrónicos, transacciones bancarias, interacciones en redes sociales, fotos, trinos, audios y sensores de vehículos y satélites, entre otros dispositivos, transmiten en forma continua data apreciable para organizaciones públicas y privadas.

"Algo muy importante ha sido la democratización de los datos. Una de las iniciativas del Centro Caoba ha sido que estas disciplinas no sean conocidas solo por personas avanzadas en matemáticas, sino por cualquiera cercano a los datos", comenta la magister en ingeniería Silvia María Lozano Argel, investigadora del Centro de Excelencia, quien trabaja en la consolidación de una comunidad de datos en el país.

Olga Lucía Quintero Montoya, doctora en Ingeniería de Sistemas de Control de la Universidad Nacional de San Juan (Argentina), fue una de las gestoras de Caoba desde el Departamento de Ciencias Matemáticas de EAFIT. Como experta en asimilación de datos e inteligencia artificial ha aplicado su conocimiento a la solución de distintos problemas.

En la actualidad, con la cooperación del Grupo de Investigación de Matemáticas y Física del Departamento de Matemática Aplicada en TU Delft (Países Bajos), apoya a estudiantes de doctorado en sus avances de vanguardia.

"El mundo real tiene dinámicas no lineales y a veces hay fenómenos que suceden que no son gaussianos –explica–. Las técnicas matemáticas aplicadas para la construcción de modelos, a



La pandemia del COVID-19 ha sido monitoreada segundo a segundo gracias a la transmisión continua de abundante data desde todos los países. Imagen con información del domingo 12 de abril de 2020 a las 11:20 a. m.



El Instituto Nacional de Salud de Colombia sigue en tiempo real la evolución de la propagación del virus COVID-19 en el territorio nacional. Imagen con información del domingo 12 de abril de 2020 a las 11:34 a. m.

partir de los datos, constituyen un área que viene desde la teoría de control y sistemas que se llama identificación de sistemas dinámicos. De ahí nació la inteligencia artificial, construir modelos que se acerquen a la manera en que los humanos razonan. Un científico de datos debe conocer la ley de los grandes números y la probabilidad para poder entender esas relaciones".

ASIMILACIÓN DE DATOS PARA CALIDAD DEL AIRE

Andrés Yarce Botero es ingeniero físico y doctor en Ingeniería Matemática

de EAFIT y Matemática Aplicada en TU Delft (Países Bajos), quien recibe la tutoría de Olga Lucía. Es experto en simplificación de datos para construir modelos basados en grafos mediante la asimilación de datos satelitales. Trabaja en un modelo matemático basado en redes libres de escala para explicar las dinámicas de tráfico de ciudades como Medellín.

Con una perspectiva de cálculos de densidades, velocidades y flujos tiene como fin construir un modelo de actividad humana que se integre al modelo de gran escala de MAUI (Medellín Air Quality Initiative), para pronosticar la

dinámica de las emisiones contaminantes. El estudio se llama *Esquemas de asimilación de datos en geodinámica colombiana: plan de investigación cooperativa para 2017-2020 entre la Universidad EAFIT y TU Delft*, en el cual apoyan la Universidad de Antioquia y la Universidad Nacional de Colombia -Sede Medellín.

"Por los problemas en calidad del aire, con la técnica de asimilación de datos buscábamos una fuente apta para ingresar a ese modelo y mejorarlo. Hay muchos sensores concentrados en ciudades de Colombia, pero en todo el territorio no hay buenas mediciones. Por eso los satélites se volvieron una muy buena fuente de datos. Utilizamos datos que vienen de satélites que miden componentes de la atmósfera como el dióxido de nitrógeno, entre otros", señala Andrés Yarce.

El estudio, aunque no se origina dentro del proyecto Caoba, usa las herramientas de la analítica avanzada en EAFIT para generar conocimiento que permite la toma de decisiones y crear soluciones para el medio ambiente. "Cuando tengamos un modelo incorporando datos de observaciones reales podemos tener la capacidad de predecir el día a día de lo que va a pasar con los contaminantes", concluye el experto.

ANÁLISIS DE REDES

"Mi tesis está fundamentada en la teoría de grafos y el análisis de redes, que pueden ser nombres diferentes para lo mismo. Lo que trato es modelar objetos del mundo real que podrían ser, por ejemplo, las características de las personas, una secuencia de ADN o una imagen digital", menciona Leandro Fabio Ariza Jiménez, investigador del Centro de Excelencia Caoba y uno de los estudiantes que hizo el doctorado en Ingeniería Matemática como parte de la Alianza.

Su trabajo, a través del uso masivo de datos, es modelar los objetos del mundo real como elementos de una red que se interconecta. Esto permite ver las relaciones entre el conjunto de datos del mismo tipo, pero representando los datos sueltos como objetos de una red. Estos grafos se basan en el concepto de entropía. "Es un asunto que es transversal porque mis datos del mundo real pueden ser muestras de ADN, estados financieros de clientes o señales del cerebro para agrupar los datos que dispara una neurona", comenta Leandro Ariza, profesor del Departamento de Ciencias Matemáticas de EAFIT.

"Una de las iniciativas de Caoba ha sido que estas disciplinas no sean conocidas solo por personas avanzadas en matemáticas, sino por cualquiera cercano a los datos".

Silvia Lozano, investigadora del Centro de Excelencia Caoba

EL NACIMIENTO DEL CENTRO DE EXCELENCIA

La Alianza Caoba surgió en 2016 como una invitación del Ministerio de Tecnologías de la Información y las Comunicaciones y del Departamento Administrativo de Ciencia, Tecnología e Innovación (Colciencias) a las empresas del sector público y privado de Colombia para fortalecer la generación de soluciones con *big data* y *data analytics*.

A ella se unieron el Grupo Bancolombia, Grupo Nutresa, IBM de Colombia, SAS Institute Colombia, EMC Information Systems Colombia, Cluster Creativ y el Departamento Nacional de Planeación. Además, participaron del acuerdo científico las universidades EAFIT, ICESI, los Andes y Pontificia Universidad Javeriana, que es el ejecutor del proyecto. Aquí se combinaron la capacidad de articulación e investigación entre academia, empresa y Estado.

EAFIT, como parte de esta alianza, acogió uno de los Centros de Excelencia para generar soluciones a problemáticas sociales, industriales y académicas que se presentan por falta de análisis masivo de información. Según el Ministerio de Tecnologías, en Colombia solo el 65 % de los datos es recolectado, de los que solo el 58,87 % es relevante. ■



Las ciencias computacionales están presentes hoy en todas las áreas del conocimiento. La nueva maestría en Biociencias de EAFIT le apuesta a la interdisciplinariedad y la investigación de sistemas biológicos complejos por medio de procedimientos informáticos. Foto Robinson Henao.